DOCUMENT RESUME

ED 133 350 TM 005 955

AUTHOR Phillips, Donald L.
TITLE Category Scoring Techniques from National Assessment: Applications to Free Response Items from Career and Occupational Development.
INSTITUTION Education Commission of the States, Denver, Colo. National Assessment of Educational Progress.
NOTE 18p.

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Elementary Secondary Education; *Essay Tests; *Guides; *National Surveys; *Occupational Tests; Reliability; *Scoring; Young Adults
IDENTIFIERS *National Assessment of Educational Progress

ABSTRACT

The Career and Occupational Development (COD) assessment of the National Assessment of Educational Progress (NAEP) was made up of about 70 percent free response exercises requiring hand scoring. This paper describes the techniques used in developing the "scoring guides" for these exercises and summarizes the results of two empirical studies of the application of these scoring guides. The guides used in the hand scoring were sets of nominal (descriptive) category systems. No attempt was made to arrange the categories along any ordinal continuum according to either quality or content. However, categories were considered to be either acceptable or unacceptable. The readers were given a scoring guide in which each category is given a descriptive title and illustrated by a number of sample responses. (RC)

NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS



# CATEGORY SCORING TECHNIQUES FROM NATIONAL ASSESSMENT: APPLICATIONS TO FREE RESPONSE ITEMS FROM CAREER AND OCCUPATIONAL DEVELOPMENT

Donald L. Phillips

National Assessment of Educational Progress

700 LINCOLN TOWER • 1860 LINCOLN STREET • DENVER, COLORADO 80203 • (303) 893-5200

# Category Scoring Techniques from National Assessment: Applications to Free Response Items from Career and Occupational Development

Donald L. Phillips
National Assessment of Educational Progress

## Introduction

The Career and Occupational Development (COD) assessment of National Assessment of Educational Progress (NAEP) was made up of about 70% free response exercises requiring hand scoring. This paper describes the techniques used in developing the "scoring guides" for these exercises and summarizes the results of two empirical studies of the application of these scoring guides.

The scoring guides used in the hand scoring of COD were sets of nominal (descriptive) category systems. No attempt was made to arrange the categories along any ordinal continuum according to either quality or content. However, categories were considered to be either acceptable or unacceptable. The readers were given a scoring guide in which each category is given a descriptive title and illustrated by a number of sample responses.

The objective of NAEP is to measure changes across time in the performance on objective-referenced exercises of the nation and several subpopulations at four ages (9, 13, 17, and Young Adults). Hence, it is important that hand scoring procedures not be dependent upon time or the context in which the item was read. Otherwise the measurement of change requires costly rescoring of responses. It is known that when essays are scored for quality on ordinal scales, the scores vary with the context in which the papers are read

(Coffman, 1971). Measurement of change in this circumstance would require all responses from all points in time to be read in the same context and time. Some NAEP research indicates that the hand scoring of COD free response items was stable across readers and some changes of context (Phillips, Burton, and Pearson, March 1975; and Phillips and Burton, November 1975).

Scoring Guides

The COD assessment used ninety-six unique items across the four age groups which were assessed. Sixty-seven of these items required some form of hand scoring. These free response items ranged from easily-scored, single answer items to difficult to score multiple answered items. Consumer mathematics items (see Attachment 1 exercise 1) are on the easier end of the continuum. An item which was a complex interview concerning jobs, job satisfaction, and other job related information represented the most difficult end of the continuum.

Most COD free response exercises had scoring guides which shared the following categories:

1. No Response
2. Other acceptable
3. Garbage
4. Other unacceptable
5. I don't know.

Only if no attempt was made to answer the question was the "no response" category used. Responses that were illegible, nonsense or obscene were coded "garbage." The "I don't know" category was

used for written responses of "I don't know," "can't remember" and other verbalized forms of I don't know. The other acceptable and other unacceptable categories were used for responses that generally met or failed to meet the general a priori standards for acceptable responses but for which there was no specific category. In addition to these conventional categories the scoring guides contained at least one descriptive acceptable category and often a descriptive unacceptable category.

Two examples of free response COD exercises and their scoring guides are in Attachment 1. Example 1 has a comparatively easy scoring guide, while example 2 has a more complex and difficult scoring guide.

NAEP develops objectives for each learning area it assesses. These objectives are generally the result of the work of a large number of learning area experts and interested laypersons acting as consultants to NAEP. Exercises are written to be measures of these objectives. The example exercises were referenced to the following COD objectives and subobjectives respectively.[1]

III. Possess skills that are generally useful in the world of work

C. Have generally useful manual-perceptual skills

4. Read displays and scales.

---

[1]See "National Assessment of Educational Progress Objectives for Career and Occupational Development," NAEP, Denver, Colorado, 1971.

I. Prepare for making career decisions

B. Know the characteristics and requirements of different careers and occupations

5. Know important factors that affect job success and satisfaction.

Hand Scoring Procedures

The hand scoring process can be roughly divided into three tasks:

1. Scoring guide development
2. Hand scorer training
3. Assessment response scoring.

In practice these three operations had considerable overlap. Some scoring guides were revised during both the training and scoring phases. These revisions made reader retraining and sometimes response rescoring necessary.

Scoring guides development started with a priori decisions as to the general kinds of response which would be considered as indicating the respondent had met the objectives of the exercise. Results from exercise tryouts on about 150 respondents were available. These results were used to help in developing categories of interest within the sets of acceptable and unacceptable responses. Some of the responses were used as examples under the category titles. Next this first set of categories was used to categorize all the tryout responses. Revisions of the categories were usually necessary to be able to accommodate all of the tryout responses.

These preliminary scoring guides were then given to the director of hand scoring. The scoring guides were tried out on some sample responses drawn from the regular assessment data. After these tryouts suggestions for further modifications to the scoring guides were sent to NAEP.*

Using these suggestions and sample responses the scoring guides underwent another revision. The scoring guides which resulted were used for the training of the readers who scored the open-ended COD items. The readers were first acquainted with the exercises, their objectives, and scoring guides. The response features which distinguished one category from another were explained. Using response data drawn from the regular assessment, all readers independently read and scored the same responses. After scoring a set the category assignments were discussed and agreements were made as to how the guide was to be used. This process continued using new sets of sample responses until the readers consistently agreed on their category assignments. For some exercises another revision of the scoring guide was necessary to get scoring consistency by the readers.

The scoring guides that emerged from the reader training sessions were usually in final form. Occasionally a number of unexpected and interesting responses of some type would occur during regular scoring. If the situation seemed to warrant it, new categories were added to the scoring guide. When this occurred,

---

*This step in the development process was possible because of a four month delay in the beginning of COD scoring. It is not part of the normal NAEP scoring procedure.

it made necessary the rescoring of already scored papers to be sure the new categories were used where appropriate.

The MRC hand scoring director checked each reader's work by rescoring a unit of their work about two to four weeks after scoring began. Where problems were detected the appropriate reader retraining and rescoring of responses took place.

At the conclusion of the scoring for each age group, the readers held a debriefing conference with NAEP's COD analyst. In these conferences the readers outlined the difficulties that had been experienced during scoring. These difficulties are being taken into account in subsequent analysis and reporting of the data.

Reliability of Scoring Procedures

Eight free response exercises from the COD assessment were used in two studies of hand-scoring stability. These studies investigated the stability of scores across time, across readers, and across age groups. In these studies the same sets of responses were read independently by every reader at two or three points in time.

In Study 1, three exercises were examined at each of the COD assessment age groups. There were seven readers involved for age 9 and either nine or ten readers at the other three ages. Twenty-eight sample responses were selected from the assessment results for each exercise at an age. Each response was read by each reader twice at age 9 and three times at the other ages. Each subsequent reading of a response was separated from the previous

reading by four to six weeks. Table 1 lists the exercises by age, the number of parts read, and the number of category assignments each reader had to make at each reading of the set of twenty-eight sample papers for each exercise.

Table 1

Study I Design

| Age | Exercise Number | (NAEP Number) | Number of Parts Analyzed (and Readers) | Number of Category Assignments (and Readings) |
|---|---|---|---|---|
| 9 | 1 | (2-301034) | 2 (7) | 56 (2) |
| | 2 | (2-302015) | 6 (7) | 168 (2) |
| | 3 | (2-402002) | 3 (7) | 84 (2) |
| 13 | 4 | (2-102025) | 3 (10) | 84 (3) |
| | 2 | | 6 (10) | 168 (3) |
| | 5 | (2-306012) | 3 (10) | 84 (3) |
| 17 | 4 | | 3 (9) | 84 (3) |
| | 6 | (2-306006) | 5 (9) | 140 (3) |
| | 5 | (2-306012) | 3 (9) | 84 (3) |
| Adult | 7 | (2-302005) | 10 (9) | 280 (3) |
| | 8 | (2-306009) | 12 (10) | 336 (3) |
| | 5 | | 3 (10) | 84 (3) |

For these two studies, the following examples illustrate how the percentages of agreement were calculated. Exercise 2 at age 13 from Study 1 is used in all three examples.

1. There were 168 separate responses to be categorized at three different times. Thus, there were 504 opportunities for all ten scorers to agree or disagree. All ten readers did agree on 461 of 504 (91.5%) possible category assignments. (See % of Time All Readers Agreed, Table 2)
2. There were 30 category assignments (ten readers on each of three readings) for each of the 168 separate responses. For each response the most frequently used category was considered the "correct" category and was called "most common score" (MCS). Each reader on each reading either did or did not give the MCS for a response. There were 5,040 (168 responses x 30 category assignments) opportunities to assign the most common category. It was assigned 5,015 times for a percentage of 99.5% (see % Agreement with MCS, Table 2).
3. There were 1,680 (168 responses x 10 readers) opportunities for a reader to agree with himself on category assignments across all three readings. There were 1,658 (98.7%) of these agreements (see % of Agreement with Self on All Readings, Table 2).

In Study I, all readers agreed upon the category assignment for a response 78.6% of the time averaged across exercises, reading and ages. Readers agreed with themselves across all readings of a response 91.5% of the time averaged across exercises and ages. The percentage of times that the most common category was assigned out of all assignments was 94.9% averaged across exercises and ages (Phillips, Burton, and Pearson, March 1975). These data appear in Table 2.

Table 2

Percents of Agreements
COD Exercises

| Age | Exercise | % of Time All Readers Agreed | % Agreement with Most Common Score | % of Agreement with Self on All Readings |
|---|---|---|---|---|
| | 1 | 93.8 | 98.4 | 98.7 |
| 9 | 2 | 74.2 | 92.6 | 90.1 |
| | 3 | 58.4 | 89.0 | 85.7 |
| | Mean | 75.5 | 93.3 | 91.5 |
| | 4 | 68.6 | 92.6 | 82.6 |
| 13 | 2 | 91.5 | 99.5 | 98.7 |
| | 5 | 72.6 | 94.0 | 86.8 |
| | Mean | 77.6 | 96.4 | 89.4 |
| | 4 | 72.6 | 93.1 | 86.9 |
| 17 | 6 | 72.9 | 93.2 | 89.4 |
| | 5 | 92.1 | 98.2 | 96.7 |
| | Mean | 79.2 | 94.8 | 91.0 |
| | 7 | 93.0 | 98.4 | 97.5 |
| Adult | 8 | 62.8 | 92.9 | 88.0 |
| | 5 | 90.1 | 96.9 | 96.6 |
| | Mean | 82.0 | 96.1 | 94.0 |
| | Grand Mean | 78.6 | 94.9 | 91.5 |

The second study analyzed an exercise given to both 9- and 13-year-olds and an exercise given to 13-year-olds, 17-year-olds, and young adults. To study age group context effect on hand-scoring, 13-year-old responses were read during the scoring of other ages for these exercises. As in Study I twenty-eight sample responses were selected for each exercise, however only 13-year-old responses were read in this study. Only those readers who read the responses for an exercise at all points in time were considered in these analyses. For the 9- and 13-year-old exercise there were four readers and there were seven readers for the other exercise. The 9- and 13-year-old exercise had six parts while the other exercise had three parts to be read. Eleven months separated the two readings for the 9- and 13-year-old exercise, about three months separated the other readings.

Table 3

Percents of Agreement on Readings of 13-year-old Responses to Two COD Exercises

| | % of Time All Readers Agreed | % of Agreement with Most Common Score | % of Agreement with Self Across All Ages |
|---|---|---|---|
| 9-13 Overlap read during 9- and 13-year-old scorings | 97.0 | 99.3 | 98.5 |
| 13-17-Adult Overlap read during 13-, 17-year-old, and Adult scorings | 83.3 | 98.2 | 95.8 |
| Mean | 90.2 | 98.8 | 97.2 |

The agreement percentages were calculated for this study like they were for Study I. All readers agreed on the category assignments 90.2% of the time averaged across both exercises. The mean percentage agreement, with most common score was 98.8% across both exercises. The readers agreed with themselves on all readings across time and age group context 97.2% of the time average over both exercises (Phillips and Burton, November 1975). These data appear in Table 3.

The percentage of all reader agreement varied considerably with the exercise from a low of 58.4% to 93.8% in Study I. More acceptable percentages for the assignment of most common score were obtained, they ranged from 89.0% to 98.4%. These data are evidence that the procedures which were used to score COD free response items provide reasonably stable data. Comparisons across time seem to be feasible as do comparisons across age groups.

## Discussion

In this section exercises from Study I will be discussed. They will be referred to by the numbers that they were assigned in both Table 1 and Table 2.

The number of categories for scoring an exercise does not seem to be related to the stability of the category assignments for that exercise. For example exercises 3 and 8, which had the lowest all reader agreement percentages (58.4% and 62.8% respectively), had relatively few scoring categories. Exercise 3 had five, six and nine categories in the guides for its parts and exercise 8 had two categories for all parts of its guide. Exercises 1 and 7

had from seven to eleven categories for each of their scoring guides, but they have two of the more stable category assignments. Exercise 6, which had sixty categories in the scoring guides for its parts, had an all reader agreement percentage near the grand mean.

There seems to be some relationship between the subjectivity of the judgments required of readers and the stability of the category assignments. The scoring guides for exercises 3, 4, and 8 were more subjective than those for 1, 2, and 7. In general the stability as reflected by the all reader agreement percentages was better for the less subjective group. The scoring guides for exercises 4 and 8 are much like that for example 2 (Attachment I), while those for exercises 1, 2, and 7 are more like that of example 1 (Attachment I). The guides for exercise 3 while unlike the others required highly subjective judgments.

As can be seen in Table I exercises 2, 4 and 5 were scored at more than one age. The guide for exercise 5 is similar to the guide for example 2 (Attachment I), but more specific. Notice on these three exercises that the all reader agreement percentages go up as we move from age 9 to age 13 and from age 13 to age 17. Age 17 and adults are fairly close on all reader agreement percentages. In general it was true that 13-year-olds gave better (more specific) answers than 9-year-olds, that 17-year-olds gave better answers than 13-year-olds and that 17-year-olds and adults gave answers of about the same quality. There appears to be a relationship between the quality of the answers and stability of the scores as indicated by the trend of all reader agreement percentages over ages. This increase in stability appears to be for both subjective and non-subjective guides.

Attachment 1

Example 1



Use the ruler you have been given to find the length of the line ABOVE. Write your answer on the answer line BELOW.

ANSWER:_____inches

SCORING GUIDE

00 = NO RESPONSE
10 = 3 3/8 OR 3 6/16
11 = ANY OTHER NUMBER FROM 3¼ to 3½
12 = OTHER ACCEPTABLE RESPONSES
20 = GARBAGE
21 = OTHER UNACCEPTABLE RESPONSES
39 = I DON'T KNOW

Example 2

Suppose a person is working on a job and has a chance to be promoted to a better paying job in which he will supervise the work of his fellow employees.

Besides the increase in salary, what are three other reasons he might accept the promotion?

(1) _____

_____

(2) _____

_____

(3) _____

_____

Scoring Guide

000 - No Response

110 - Wants a challenge, responsibility or wants to be more useful; learn new things.

A Change to me more responsible.
Like the challenge of something new.
If you thought you could help people under you.
A chance to put into work some of your own ideas.
Wants a harder job.
Learn more about people.

111 - Personal satisfaction with new job and duties.

Satisfaction of seeing the job well done.
Could like the work better than his previous job.
He would feel an achievement of being promoted.
Better type of job.

112 - Prestige, status, being the boss, more power.

He will be supervisor.
Higher authority.
Enjoy being in charge.
He could tell his employees what to do.
Most people want to get up in the world.

113 - Open opportunities for further advancement.

He will have a chance to get more promotions.
Higher in rank - might be promoted again later.
You think it will help your future.

114 - not used

115 - Better working conditons, fringe benefits and other mechanical aspects.

Not as much manual labor.
Get days off she wanted.
Less overtime.
Better retirement benefits.

116 - Better location.

It might be closer to home than his other one.
He might want to change his surroundings.

117- Feels deserving or capable.

He has worked there a long time and he deserves it.
He has the ability and know how to supervise.
He would feel he is the best one to take the job.

118 - Interpersonal relations.

He will still work with his fellow employees.

119 - Other acceptable.

Family gratification

250 - garbage

251 - Other unacceptable

To earn it.

Comparison of other work.

May be like people.

Its a better job.

252 - Exact duplicate of a previous response.

253 - Salary

For the money.

He could use more money to buy more things.

The high cost of living.

399 - I don't know

## References

Coffman, William E. "Essay Examinations" in Educational Measurement, Robert L. Thorndike, Editor. American Council on Education, 1971. Washington, D.C.

Phillips, D.L., Burton, N.W., and Pearson, A.M. "Stability of Nominal Categories Over Readers, Over Time," National Assessment of Educational Progress, March 1975. Denver, Colorado.

Phillips, D.L., Burton, N.W. "Stability of Nominal Categories Over Ages," National Assessment of Educational Progress, November 1975. Denver, Colorado.